# 25th ANNUAL NATIONAL CONFERENCE ON MANAGING ENVIRONMENTAL QUALITY SYSTEMS

## APRIL 24-27, 2006

## Marriott Renaissance, Austin, Texas

## Technical Papers

Electronic Tools to Improve Data Management

- Training a Prototype Classification Algorithm to ID Contaminents for the CCL List (M. Messner)
- A TRI Trend Analysis Tool (N.Neerchal)
- Utilization of 3-D Spinning Plots as Exploratory Tool s (H.Allendar)
- New Environmental Stats Software GiSd (Guided Interactive Statistical Decision Tools ) (D. Bronson, K. Black)
- HMSI Monitoring System: A Tool for Long-Term Remediation Performance Monitoring (W. Michael)

# Training a Prototype Classification Algorithm to Identify Contaminants for the Contaminant Candidate List (CCL)

Michael Messner, Thomas Carpenter, and Zeno Bain
USEPA Office of Ground Water and Drinking Water
Washington, DC 20460

Abstract

*In the past, EPA experts would periodically consider relatively small numbers of chemical and microbial contaminants for inclusion in EPA's Contaminant Candidate List. Recently, a National Research Council committee and a National Drinking Water Advisory Council recommended the use of computer-based methods to facilitate consideration of a broad "universe" of contaminants.*

*This paper describes recent work by EPA's CCL Team to:*

- *Define and score the attributes of chemical contaminants*
- *Select computer algorithms*
- *Develop a training data set*
- *Express concerns for the different kinds of classification error*
- *Training the classification algorithms*
- *Visualize and assess the algorithm outputs*

## Scoring the Chemical Attributes (potency, severity, prevalence, and magnitude)

Attribute scoring protocols were developed to (a) communicate the degree to which a scored contaminant exhibited the attribute, while (b) providing some separation, or distinction between contaminants. Because potency, prevalence, and magnitude tended to span many orders of magnitude, these scoring protocols tended to be logarithmic.

Each protocol identifies a contaminant score based on the strongest available type of data for the contaminant. For example, a contaminant having results from analysis of numerous drinking water sources (e.g., surface waters) would be scored on those data, rather than on production and use data, because direct source water data is the more reliable for predicting the amounts that would appear in drinking water.

Protocols were adjusted iteratively, as the training data set was developed and discrepancies were observed between blinded, unblinded, and rule-based classifications.

## Selecting Classification Algorithms / Models

Five classification methods were considered: Artificial Neural Networks (ANN), Classification and Regression Tree (CART®), Multivariate Adaptive Regression Splines (MARS®), simple linear models, and QUEST® (Quick, Unbiased, Efficient Statistical Tree). The five methods / programs are all very quick and easy to use, but differences in transparency and quality appear to be significant. The table below lists the features of the three most favorable methods.

| **Features** | **Artificial Neural Network** | **Classification Tree with Linear Nodes (QUEST)** | **Linear Regression** |
|---|---|---|---|
| **Objective Function (to be minimized or maximized)** | Minimize count of training set errors | Minimize count of training set error loss OR minimize error loss | Maximize likelihood or minimize error loss |
| **Prediction** | Rounded average team classification | Rounded average team classification | Average team classification (not rounded) |
| **Ranking Capability** | Rank by Pr(List) | Rank by classification and distance from discriminant (requires post-processing) | Rank by prediction |
| **Transparency of Optimization Method** | Not transparent | Not transparent | Simple and transparent |
| **Classification Rule** | Not clear, but classifications available for all attribute score combinations. | Clear. Classification tree with linear inequalities for intermediate nodes | Clear. Simple linear function of attribute scores. |
| **Computation Speed** | < 1 Second | < 1 Second (but process for deriving distances is not automated) | < 1 Second |
| **Software Cost** | ??? | Freeware | No special software |

## Developing the Training Data Set

Initially, the training set consisted of contaminants that were readily available, with strong occurrence and health effects information. These included contaminants currently regulated in drinking water, contaminants included in earlier Contaminant

Candidate Lists, and contaminants "generally regarded as safe," a designation given by the U.S. Food and Drug Administration.

CCL team members first attempted to classify these contaminants as either "List" or "Not List," according to all of the occurrence and health effects information available. Having difficulty with the two options, team members were more comfortable expressing their judgments when two less certain categories were added: "Not List?" for contaminants that seemed to belong off the CCL, but with some uncertainty, and "List?" for contaminants that seemed to belong on the CCL, with some uncertainty. After assigning contaminants to the four categories (designated NL, NL?, L? and L), team members discussed their assignments and were allowed to make adjustments based on what they learned in the discussions.

As scoring protocols developed, the team considered the same contaminants, but viewing only their integer attribute scores. Contaminant names and supporting data were not revealed. Again, team members discussed their assignments and adjusted accordingly. When contaminant names were finally revealed, the information-based assignments were compared to the new "blinded" score-based assignments. When the blinded and unblinded assignments differed, the team generally agreed that the unblinded assignments were the more appropriate for training purposes. These differences provided the basis for iterative adjustment of the scoring protocols.

The first 102 contaminants were real chemical contaminants. This training set was supplemented by addition of artificial contaminants. The majority of these were selected using Latin hypercube sampling from the set of all possible attribute score combinations. A small number were deliberately selected to fill in some obvious voids in the 4-attribute space.

Team-based classifications were found by averaging the classifications of team members and rounding to the nearest integer. Ties (such as having two members select List = 4 and two select List? = 3 resulting in an average of 3.5) were rounded to the higher integer (3.5 rounded to 4, List). ANN, CART, MARS, and QUEST utilized these rounded assignments, while the linear model utilized the raw team averages so it could estimate the average as a linear function of attribute scores.

## Expressing Concerns for Different Kinds of Classification Error

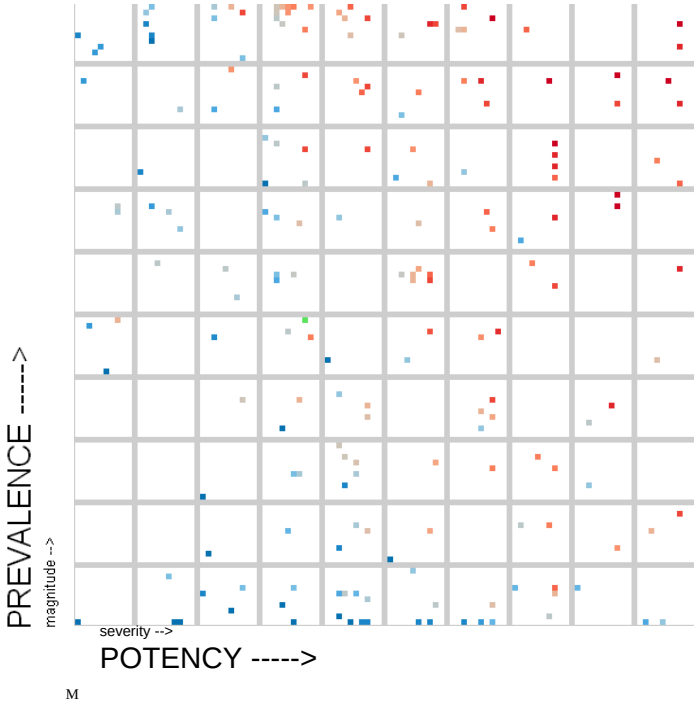
The table below shows the errors that are possible, given the four-category problem. Of these, the most serious would be placing a strong "List" contaminant in the "Not List" category. After applying the algorithm-based rules, the Team plans to scrutinize all of the contaminants identified as "List," only a fraction of those identified in "List?" and very few of those placed in the other categories. As a result, we recognized the need to minimize the likelihood of classifying strong contaminants as "Not List" or "Not List?" In contrast, we expect to scrutinize every contaminant assigned to the "List" category, identifying those errors, so their only costs would be the time and effort required for our Team to review the data and check them.

Considering the relative seriousness of the different kinds of errors, the Team represented the error losses in terms of the weights displayed in the table. The most serious error (placing a List contaminant in the Not List category) has ten times the cost of the least serious error (placing a contaminant one category too high, such as placing a List? Contaminant in the List category).

| | | Correct Assignment (Team) | | | |
|---|---|---|---|---|---|
| Algorithm-Based Rule's Assignment | Assignment | NL = 1 | NL? = 2 | L? = 3 | L = 4 |
| | NL | ---- | 2 | 5 | 10 |
| | NL? | 1 | ---- | 2 | 5 |
| | L? | 2 | 1 | ---- | 2 |
| | L | 3 | 2 | 1 | ---- |

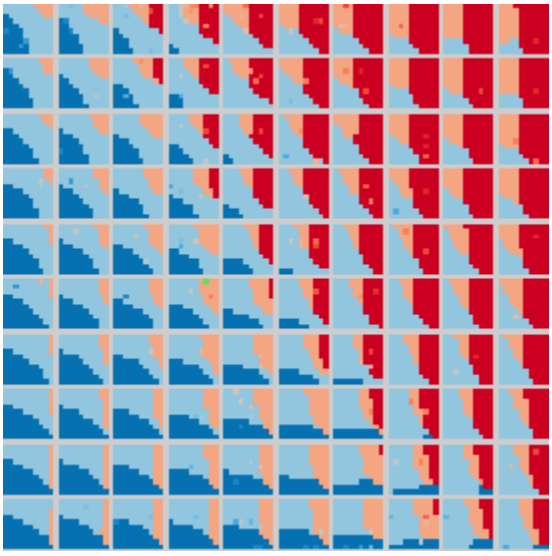
## Training the Classification Algorithms

Training was done on a number of occasions: when batches of new contaminants were added to the training data set, or when changes were made to the attribute scoring protocol (therefore necessitating a change to the contaminants' scores). There are currently 202 contaminants in the training data set. The figure below shows their placement in terms of attribute scores. Colors reflect the average team classification, unanimous List = red to unanimous Not List = dark blue. One contaminant (Potency = 4, Severity = 8, Prevalence = 5, and Magnitude = 10) is shown in green, though the team's decision for that contaminant is List. This particular contaminant is always shown in contrasting color to ensure that the axis labels are correct. We also have the ability to exchange axes so that contaminants and rules can be viewed from other perspectives.

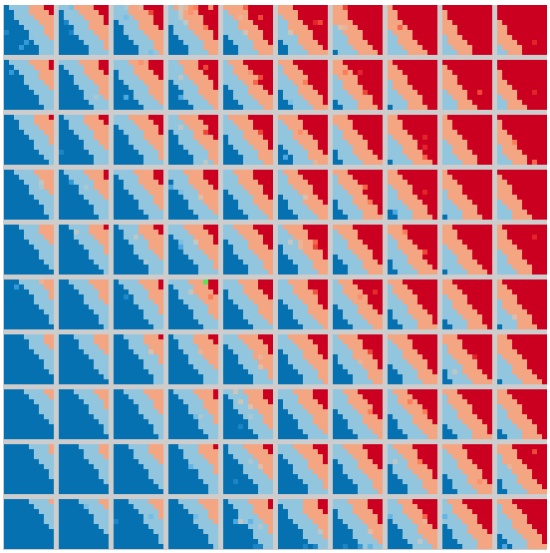


**Visualizing and Assessing the Algorithm Outputs and Performance**

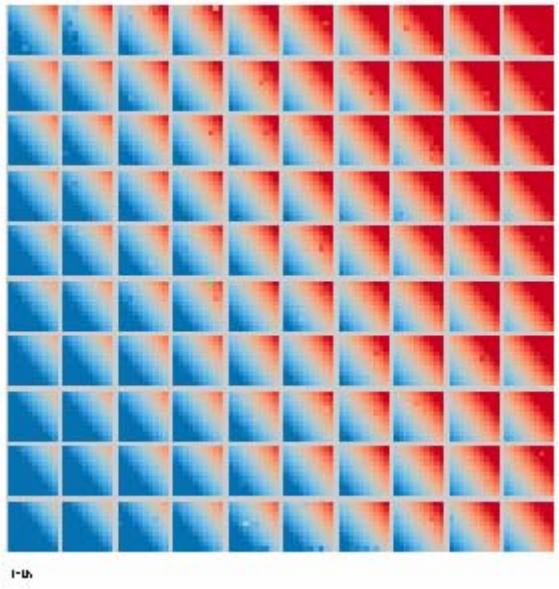
The figure below illustrates a problem with the MARS-based rule. Notice that the training set contaminants appear as small points. The orientation point (which would otherwise be red) appears again as green. Notice that it is surrounded by the color for List? The figure shows areas where red touches light blue and where dark blue touches beige. Both are indications that the algorithm was unable to define the intermediary categories.



Below, the results of the ANN algorithm show no such inconsistencies. Red never touches blue and dark blue never touches beige. The training data points seem to agree better with the ANN rule: they are less "obvious."



The linear model predicts average team classification of contaminants. Average classification is not necessarily an integer, so actual and estimated classifications can be shown in intermediate colors, as shown below.



Based on training with the current data set, the CART and MARS algorithms exhibited inappropriate behavior, while the other three approaches (ANN, linear model, and QUEST) performed very well with respect to training set error loss and number of training set errors. The linear model was generally able to predict the team average within approximately 0.3. Although decisions have not been made about discarding CART and MARS, the Team is working through some procedures for combining the results from the other three approaches.

## A TRI TREND ANALYSIS TOOL

*Justin Newcomer, Nagaraj K. Neerchal, Pepi Lacayo, Barry Nussbaum*
*University of Maryland Baltimore County & OEI*

TRI Explorer is a web-based tool which can be used by an analyst to download slices of the TRI data base. While the tool works very fast, extracting data needed to perform comparative study of trends for a group of chemicals may need to be done using several downloads from TRI Explorer and then some manipulation of the data. We provide a simple offline tool which provides more customized options to analysts for slicing and dicing the TRI database. One of the key features of the tool is its ability to extract multiple years of TRI data, simultaneously to create profiles for estimating trends. The tool is also expandable to include more data, as they become available on annual basis.

# Utilization of 3-D Spinning Plots as Exploratory Tool for Pesticide Residue Changes throughout the Years

**Hans D. Allender, PhD, PE**

**allender.hans@epa.gov**

## Introduction

Regulatory agencies, as well as industry, are in a constant search for ways to demonstrate efficacy in what they do; these organizations produce and present data to their customers and stake holders to inform then about the outlook of the business and allow them to produce better decisions going forward. Classic graphic displays (i.e., pie charts, bar charts, line charts, etc.) are abundant in every presentation that measures and compares the state of the business. The objective of this paper is to explore the powerful use of 3-D spinning plots to develop indicators. This tri-dimensional technique may help the audience to understand key concepts and to visualize data in an influential way not experienced before.

## The Spinning Plot

Spinning plots were introduced by Fisherkeller et al. circa1974. Today with the development of fast computers and specialized software, spinning plots can be constructed efficiently. To build the spinning plots, this presentation uses the JMP statistical discovery software. JMP is a product of the SAS Institute and brings to the user a complete line of general statistics and graphical representations in a point and click fashion. The plot is a spinnable display of the values of numeric columns in the current data table. The Spinning Plot platform displays three variables at a time from the columns you select in the data table. The **Spinning Plot** command in the **Graph** menu (or toolbar) displays a three-dimensional view of data and an approximation of higher dimensions through principal components. You can also launch the Spinning Plot platform with the **Spinning Plot** button on the **Graphs** tab page of the JMP Starter window. To help capture and visualize variation in higher dimensions, the spinning platform displays a bi-plot representation of the points and variables when you request principal components.

## Spinning Plots: an Application

In this particular application, the objective is to visualize changes of pesticide residue on fruit and vegetables throughout the years using the Pesticide Data

Program (PDP) database. The U.S. Department of Agriculture implemented PDP in 1991. Since then, PDP has tested a wide range of commodities in the U.S. food supply. Using a rigorous statistical approach and the most current laboratory methods, PDP has tested both fresh and processed fruit and vegetables, grain, milk, beef, and poultry. PDP data are essential for the implementation of the 1996 Food Quality Protection Act, which directs the Secretary of Agriculture to collect pesticide residue data on foods most likely consumed by infants and children. The EPA uses PDP data as a critical component of pesticide dietary assessments.

**What is in a point?**

In order to represent the values of the data base in a tri-dimensional space we need to select measurements that identify typical information on the crop-pesticide pair (for example carbaryl on apples). In this study the parameters selected were

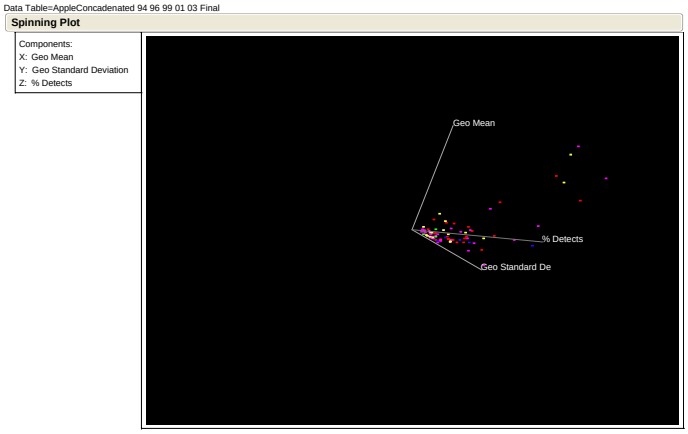
- The Geometric Mean
- Geometric Standard Deviation
- % Detects

The geometric mean produces a measurement of central tendency of the sample crop-pesticide pair for a particular year, the geometric standard deviation tells us about the variability of the sample, and the % of detects indicates the proportion of residues detected on the crop. When the crop is analyzed in the laboratory, the pesticide residues may be zero or an amount so small that the lab equipment can't detect it. A large fraction of detects implies that the pesticide is prevalent in the given crop.

These three statistics are calculated for each year that data is available, and are represented as a point in the tri-dimensional space for each combination of crop, pesticide and year. To complete the identification process it's necessary to color-code each year. Equipped with these conventions, we are ready to display the spinning plots and hopefully to discover trends that take the points closer and closer to the origin. In other words, the closer the points get to zero (the origin) the better the public is served because this implies a reduction of the geometric mean, a reduction of the geo standard deviation and a reduction of the % of detects; this is to say, a reduction of pesticide residues on food.

## Examples of Spinning Plots

# Pesticides in Apples (94,96,99,01, and 03)




**Points**

**1994**

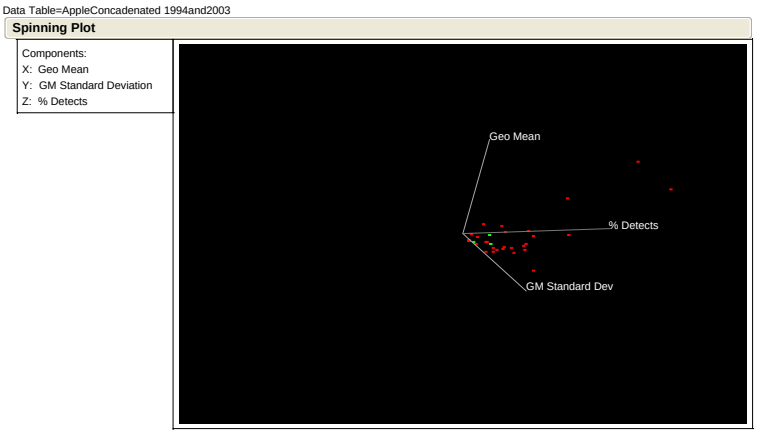
**1996**

**1999**

**2001**

**2003**

# Pesticides in Apples (94 and 03)




**Points**

**1994 (24)**

**2003 (3)**

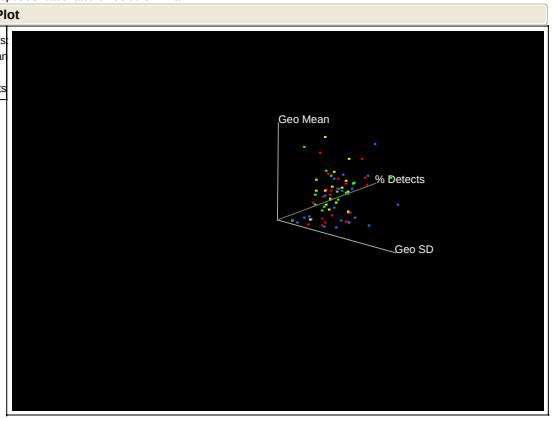
**Reduction 87%**

# Pesticides in Grapes (94,96,00,01)

Data Table=GrapesConcadenated 94 96 00 01 Final

Spinning Plot

Components
X: Geo Mean
Y: Geo SD
Z: % Detects




**Points**

**1994**

**1996**

**2000**

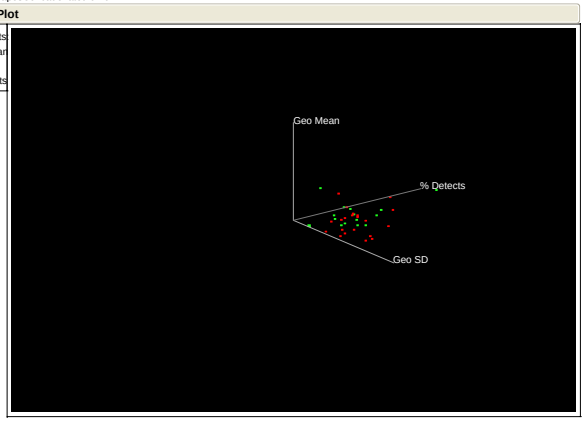
**2001**

# Pesticides in Grapes (94 and 01)

Data Table=GrapesConcadenated 94 01

Spinning Plot

Components
X: Geo Mean
Y: Geo SD
Z: % Detects




**Points**

**1994 (21)**

**2001 (15)**

**Reduction of 28%**

In this live demonstration, the presenter will show spinning plots that allow viewing different angles of the space composed by Geo mean, Geo standard deviation, and % crop treated. These angles permit tracking of specific pesticides through time and isolation of specific chemicals for further study.

**Conclusions**

- The spinning plots can provide a valuable visual tool to determine and explain trends.
- Geometric means, geometric standard deviation, and % detects are three statistics that provide valuable summary information on the distributions of pesticide residues on agricultural crops by year.

# NEW ENVIRONMENTAL STATISTICAL SOFTWARE: GiSdT

Doug Bronson (dbronson@neptuneinc.org)
Kelly Black (kblack@neptuneinc.org)
Neptune and Company, Inc.

## Abstract

*While there are many statistical software packages available, most have a steep learning curve and require an underlying knowledge of statistical concepts. Neptune and Company has designed and built a free, web-based program that will help users get the gist of statistical concepts without requiring them to learn a programming language or to fully understand the underlying theory. Our software, "GiSdT", is available on-line at www.gisdt.org. GiSdT includes many statistical methods described in EPA Quality Staff's G-9S Guidance on Data Quality Assessment. By providing guidance on the pros and cons of these statistical methods, GiSdT assists the user in making reasonable choices for analyses, and leads to efficient, statistically supported decision-making without requiring the user to purchase or learn complex statistical software.*

## Introduction


GiSdT (www.gisdt.org) is an open-source, web-based, decision support system for data analysis. GiSdT combines the power of the Internet with analysis and presentation tools that can be used interactively to solve statistical problems. The objective of GiSdT is to provide an interactive technical guidance program with analysis capabilities that is developed with open source software. While the framework for GiSdT is in place and many analysis tools are available, other tools and capabilities still need to be developed. The advantages of GiSdT include:

- Open-source, web-based tools so users aren't encumbered with software licensing issues.
- Allows users to engage at various levels of complexity depending upon their interest. For example, different levels of complexity might include (for each component): an overview, access to supporting information, specific analysis, details of mathematical methods, and access to computer code, e.g., XML (Extensible Markup Language) or R (an open source statistical programming language – www.r-project.org).
- An overall approach that facilitates defensibility, traceability, and transparency. The web-based tools that are used are completely open for inspection and review. There is no proprietary code and the source code for every component of GiSdT is presented and available.
- GiSdT is a living web-based system that can be updated as new tools, technologies, and approaches become available.
- GiSdT is software independent; the content can be edited in any text editor. The GiSdT content can be continually built upon. The presentation of the GiSdT content can be dynamic and tailored to best meet the needs of the user community.
- Does not require knowledge of a statistical programming language.
- Features a personal user project space where user data files and user created results files are stored for easy recall.

## Typical Data Analysis Session

To begin analysis, the user uploads one or more datasets to a personal GiSdT project space called "My Project", as shown in Figure 1. Then a method is selected (classical statistical methods currently available in GiSdT are shown in Table 1) and input parameters chosen via a simple user interface like the one in Figure 2. After submitting the desired input parameters, topic results are returned to the screen and stored in "My Project". Figure 3 shows an example project space. This project interface allows users easy access to their datasets and any analyses they have performed. Finally, Figure 4 shows example output from GiSdT. Output can be easily exported to any word processing or spreadsheet program.

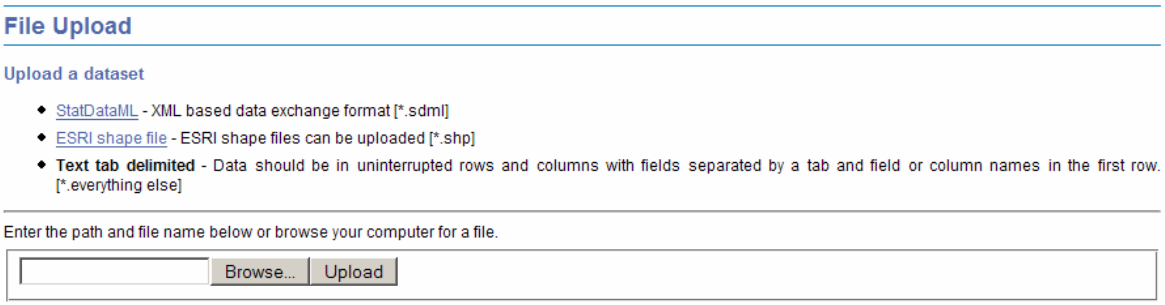



**Figure 1. Interface for Uploading Data**

| Exploratory Data Analysis | One-Sample Methods | Two-Sample Methods |
|---|---|---|
| Summary Statistics<br>Boxplots<br>Histograms<br>Normal QQ-plots<br>Bubble/Intensity Plots | t-test<br>Sign Test<br>Wilcoxon Signed Rank Test<br>Confidence Intervals/Limits<br>Tolerance Limits | t-test<br>Wilcoxon Rank Sum Test<br>Quantile Test<br>Slippage Test<br>Sign Test<br>Wilcoxon Signed Rank Test<br>Comparison to Background |

**Table 1. Classical Statistics Methods Available in GiSdT**

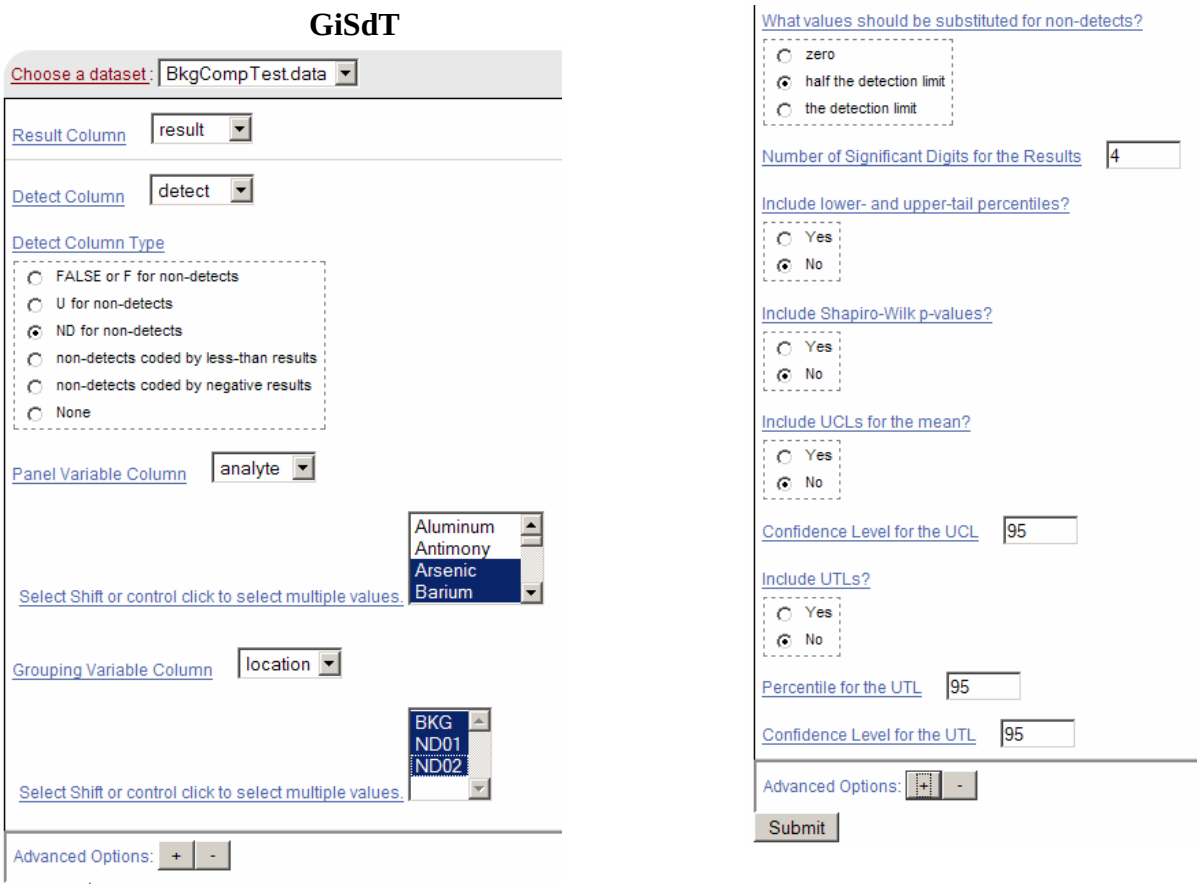



**Figure 2. Example GiSdT User Interface**

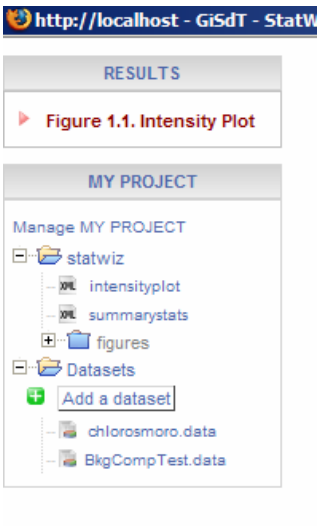



**Figure 3. Example Personal GiSdT Project Space**

GiSdT - StatWiz - Mozilla Firefox

## Results

Next
[Page 1 of 2]

### Summary Statistics

Summary Statistics

| analyte | location | N | Num Detect | Min ND | Max ND | Min Detect | Median | Mean | Max Detect | Std Dev |
|---|---|---|---|---|---|---|---|---|---|---|
| Arsenic | BKG | 45 | 39 | 1 | 1 | 5.2 | 11.1 | 20.98 | 107 | 22.73 |
| Arsenic | ND01 | 58 | 58 | NA | NA | 2.3 | 8.85 | 10.57 | 28.8 | 6.395 |
| Arsenic | ND02 | 77 | 75 | 0.3 | 0.3 | 2.3 | 10.2 | 11.71 | 35.7 | 6.718 |
| Barium | BKG | 45 | 27 | 10 | 20 | 49.6 | 61 | 94.81 | 839 | 140 |
| Barium | ND01 | 58 | 58 | NA | NA | 54.6 | 207 | 235.4 | 770 | 133.4 |
| Barium | ND02 | 77 | 69 | 10 | 20 | 44 | 96.4 | 107.4 | 596 | 77.76 |

Next
[Page 1 of 2]
Top

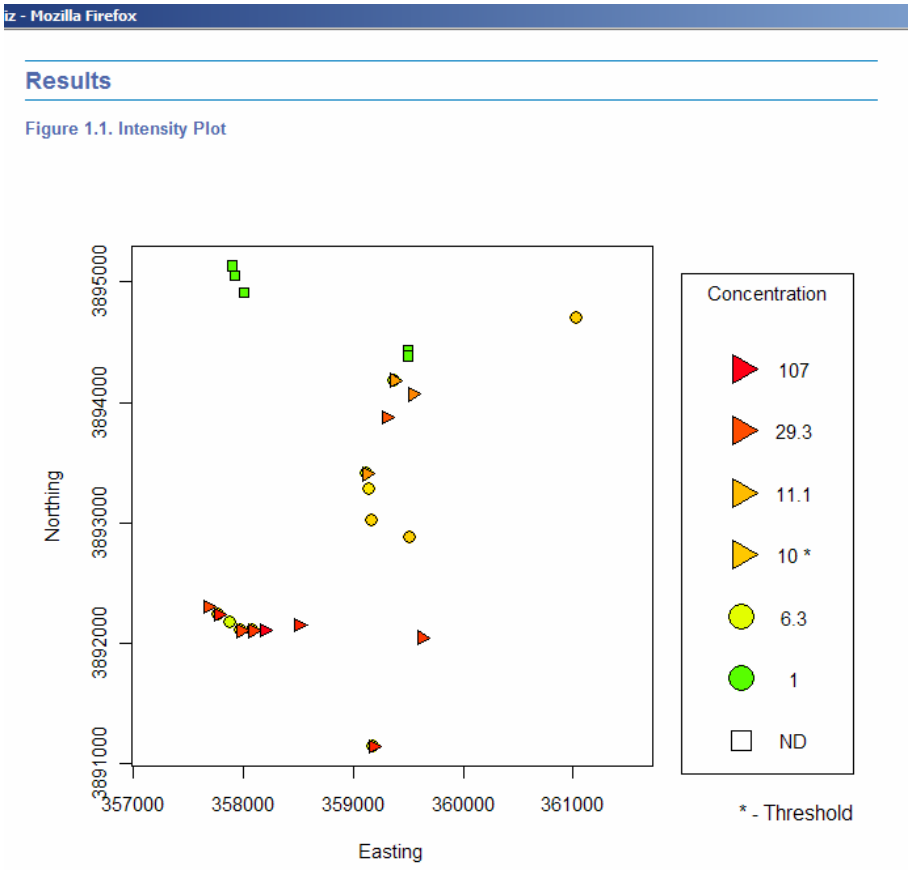



**Figure 4. Example Output From GiSdT**

# THE HMSI MONITORING SYSTEM: A TOOL FOR OPTIMAL AND EFFECTIVE LONG-TERM REMEDIATION PERFORMANCE MONITORING

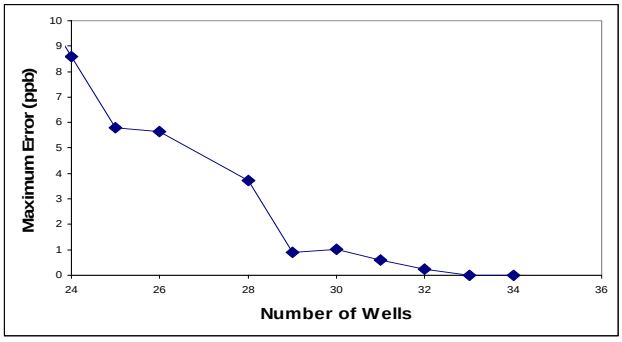
*Bill Michael[1], Barbara Minsker[1], Matt Zavislak[1], Charles Davis[2] and David Tcheng[3]*
*[1]Hazard Management Systems Inc*
*[2]Environmetrics & Statistics Limited*
*[3]National Center for Supercomputer Applications*

Many organizations in both private and public sectors incur significant costs associated with performance monitoring of subsurface remediation. This presentation presents a demonstration and evaluation of the HMSI Monitoring System, a new tool for system optimization, troubleshooting, and proving progress at an Atlantic Richfield remediation site in New Jersey. The HMSI Monitoring System consists of three integrated components:

Model Builder creates geostatistical or analytical models of spatial and/or temporal trends based on historical data. It supports both automated and manual model-fitting approaches.

System Optimizer identifies sampling location and/or frequency adjustments that would be most beneficial for continuing monitoring based on redundancies in sampling locations in past data. This analysis uses multi-objective genetic algorithms, allowing users to readily identify optimal monitoring tradeoffs; the figure below shows an example. The Optimizer can perform simultaneous temporal and spatial analysis

Data Tracker allows users to specify monitoring targets that are both based on historical patterns and consistent with site-wide data quality objectives. These targets can include estimated contaminant mass reductions or contaminant level goals at key locations. Data Tracker can then be used to automatically evaluate new data and identify significant deviations from the targets. Alerts can be issued to notify users where their attention is most needed. Data visualizations (e.g., interpolated plume concentrations over time) assist both in troubleshooting and in proving remediation progress.




This presentation will introduce these components, report on their effectiveness at improving performance monitoring at the demonstration site, and describe the success criteria for future applications.

This figure presents the results of a spatial redundancy analysis for groundwater monitoring at the demonstration site. Each point represents an optimal monitoring strategy for a given number of monitoring wells. The Monitoring System identifies optimal monitoring well networks as a function of cost, allowing the user to decide among tradeoffs between cost and spatial and temporal interpolation errors.